Assignment-based Subjective Questions

Question 1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer-The optimal value of alpha of Ridge=20, Lasso=500. Let us double the values of alpha for Ridge=40 ,Lasso=1000. We observe-

a. **Slight decrease in R2 score and increase in other error metrics in both Training and Test set is observed. Metrics shown in below table.**

Metrics	Ridge(alpha=20)	Ridge(alpha=40)	Lasso(alpha=500)	Lasso(alpha=1000)
R2 score(Train)	0.884	0.873	0.855	0.828
R2 score(Test)	0.869	0.867	0.851	0.831
RSS(Train)	735114863758	806413605286	920060502527	1091509915696
RSS(Test)	367938226113	374390386197	418702737905	473802130290
RMSE(Train)	26832	28103	30018	32696
RMSE(Test)	28983	29236	30918	32889

b. **Most important predictor variables after doubling alpha is following—**

For Ridge

	Features	Coefficients
0	MSSubClass	173485.934734
4	OverallCond	17035.035748

	Features	Coefficients
68	Neighborhood_NridgHt	16604.843881
69	Neighborhood_OldTown	14172.003328
16	BsmtFullBath	13960.826882
169	BsmtExposure_Mn	11991.803108
14	LowQualFinSF	11582.501423
79	Condition1_PosA	9927.346708
26	GarageArea	9471.741830
59	Neighborhood_Edwards	9406.161837

For Lasso

	Features	Coefficients
0	MSSubClass	182977.419591
16	BsmtFullBath	25386.130770
4	OverallCond	23403.502850
69	Neighborhood_OldTown	11307.027295
26	GarageArea	10234.283703
6	YearRemodAdd	8150.885689
68	Neighborhood_NridgHt	7713.698984
79	Condition1_PosA	7653.997334

	Features	Coefficients
169	BsmtExposure_Mn	6174.992344
17	BsmtHalfBath	5184.566798

For Ridge- In top 10 features, Feature -Exterior1st_AsphShn is replaced by GarageArea but rest features still remain in top 10,also order of top 10 features is slightly rearranged.

For Lasso -In top 10 features, Feature -Neighborhood_Edwards is replaced by BsmtHalfBath but rest features still remain in top 10,also order of top 10 features is slightly rearranged.

Question 2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Answer-

a. **The optimal value of lambda for Ridge=20, Lasso=500.**

b. **In the Test set, the R2 score of Ridge is 0.86 vs. R2 score of Lasso which is 0.85 which are almost similar. Lasso has an advantage over Ridge, that it helps in feature reduction by setting many of the coefficients to zero and hence prevent overfitting. Hence our choice will be Lasso.**

c. **We practically see no overfitting in our Lasso model. The R2 score for Lasso Regression for both Training and Test is 0.85 and hence clearly there is no overfitting, since the model performs equally well on both Training and Test set.**

Question 3

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer-

a. **Top 5 features initially were MSSubClass, Neighborhood_NridgHt, BsmtFullBath,**

Neighborhood_OldTown, OverallCond. Let us rebuild the model without top five predictor variables.

b. **After excluding the top five features, the new top 5 features are -LotFrontage, Neighborhood_SWISU, BsmtHalfBath, YearBuilt, BsmtExposure_Mn.**

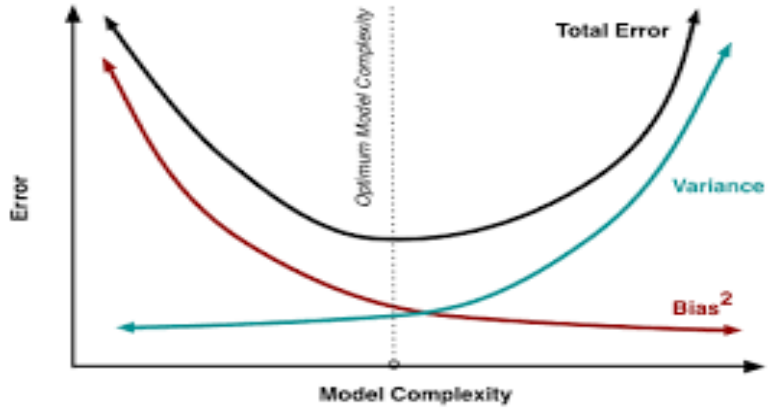
c. **After dropping top five features we observe a drop in R2 score from 0.85 to 0.84.**

d. **We observe an increase in RMSE on Test set from 30918 to 31535.**

Question 4:How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer- To make sure the model is robust and generalizable we have to understand the bias-variance tradeoff. A high bias means the model is weak learner and is unable to pick fine details and patterns in the data. A very simple model suffers from high bias and performs poorly on training and test data. A complex model performs very well on training data as it overlearns the fine details but fails to generalize it. In case of test set or slight changes in data , the model fails to generalize and performs really bad. A very complex model suffers from high variance.

To make sure the model is robust and generalizable we have to find the sweet spot between bias and variance so that we do not suffer from under/over-fitting as shown.



To make models robust and generalizable, techniques such as regularization are used. Ridge and Lasso regression are popular choice for the same. In Ridge and Lasso we have to choose the optimal value of lambda so that the penalty term is optimal(this would help determine the model complexity). This ensure the model is of right complexity and only the relevant features are used to build the model so that it does not become overly complex(for example Lasso sets many coefficients to zero).

The implications of this on accuracy is that a robust and generalizable model will perform equally well on training and test data. This was seen in the Ridge and Lasso models we built. A non-robust and non-generalizable model may perform better than robust, generalizable

model on training set because of overfitting and high complexity but accuracy on training and test set will vary vastly and performance on test set will be bad.